TERMINAE User Manual - V11-2

Sylvie Szulman (Paris 13)

with contributions from: Adeline Nazarenko (Paris 13)

2011 July

Abstract

TERMINAE is a platform that assists users in designing termino-ontological resources from texts. It can be used by terminologists to build terminological forms and by knowledge engineers to build either thesaurus expressed in SKOS or ontologies organising concepts and lexical units in a formal way supporting inferences.

This platform allows to link textual elements to terminological and conceptual resources. The acquisition corpus may contain one or several documents. The supported languages are English and French.

Keyword list: Ontology acquisition, terminology, assisting tool

Executive Summary

This document is the user guide of TERMINAE.

TERMINAE is a platform that assists users in the design of termino-ontological resources from texts. In ONTORULE, it is used to build from texts

- thesaurus expressed in SKOS, and
- ontologies organising in a formal way the concepts associated to the terms and supporting inferences.

This platform allows to link textual elements to terminological and conceptual resources. The corpus may contain one or several documents. The supported languages are English and French.

TERMINAE is organised in three main levels: the first step of the terminological level enables to constitute the set of terms of the corpus; its second step organises these according to lexical and syntactic relations; the termino-conceptual level organizes the terminology according to semantic relations; the third level, the ontological level, enables to create a formal ontology out of the list of termino-concepts created at the second level.

This document describes the functionalities of the TERMINAE platform. The first chapter describes the technical characteristics and the installation instructions. The following chapters present the main menus of the platform that are accessible from its main window.

Contents

Chapter 1

Introduction

This document describes the functionalities of the TERMINAE platform which is an eclipse application. Chapter 2 gives a very short insight of the methodology. Chapter 3 gives the technical characteristics and the installation instructions. Chapter 4 presents the main menu and the following chapters (chapters 5 to 9) introduces the 5 perspectives of the platform and the related functionalities.

Chapter 2

The Terminae method

TERMINAE is a tool that is supported by a method, and some (very short) forewords on the method can help using the tool. The task is to build a domain ontology. This is an expert task, since it needs to decide which concepts are really important for the domains, and how they are related. It has been experienced that linguistic tools, relying on texts specific of the domain, can help the expert. They do not do the work in his/her place, but they propose a good starting point to improve the coverage of the domain, and some ambiguities they raise reveal real and unseen ambiguities of the domain vocabulary.

The TERMINAE method starts from the linguistic results produced by TreeTagger and YaTeA. It has then three steps.

- At the linguistic level, the input is a list of **term candidates**, i.e. words or group of words which, on a linguistic basis, could possibly figure in a **terminology** of the domain (a list of its main terms). The goal of this level is in a first step (chapter 6 to constitute, clean and improve the list, removing parasistic or irrelevant proposals. A second step (7) involves grouping those which are morphologic variants of the same term and collecting linguistic relations. This work relies on the list of occurences of each term, which are gathered with linguistic information in **terminological forms**.
- The **termino-conceptual** level (chapter 8) is specific to TERMINAE. Whereas terms are at the vocabulary level, the goal is now to analyse the use of terms in the corpus at the semantic level. The work is to recognize and distribute the various senses of this term into several **termino-concepts**, distributing also the occurences of the term between senses. At the same time, the termino-concepts of the form can be tagged as having a synonym in an other form, or being otherwise (more loosely) related.

- The **ontological** level (see chapter 9) now relies on termino-concepts and their relations to build the ontology. First, synonym termino-concepts should only yield one concept. All the related termino-concepts help building the hierarchical relations and defining the roles, as can do some other linguistic information gathered during the process - a part of which is under exploration in the framework of the ontorule project (in particular the analysis of verbs and SBVR fact forms).

Chapter 3

Technical Characteristics

- The current version of TERMINAE platform is compiled using SUN 1.6 Java virtual machine.
- It relies on UTF-8 text encoding.
- It can be used for English and French.

3.1 Installation

To install TERMINAE, you need java, version 1.6. Download the version of the platform for your system from the

`http://www-lipn.univ-paris13.fr/~szulman/TerminaeWorkbench/`

web page and unzip the downloaded file.

The default language is English but it can be changed. If you want to work with a French platform, edit the `terminae.ini` file and change the line `nl en` by `-nl fr_FR`. This file is located in the `Terminae` directory on Linux and Windows systems and in the `Terminae.app/Contents/MacOS` directory on MacOS systems.

3.2 How to start

To launch the TERMINAE platform, click on the `Terminae` application (either `Terminae` on Linux system, `Terminae.exe` on Windows system or `Terminae.app` on MacOS).

Initially, the project management perspective (`Terminae Project perspective`) is open and you have to import or create a project.

3.2.1 Project location and structure

In any case, you have to define your project directory. On Linux and Windows systems, it is advised to locate it in the `workspace` directory created by the eclipse application.

A project has a fixed structure, represented as the 6 following subdirectories:

- `corpora`: Contains the corpus data (raw and tagged) and the results of named entity recognition tools. The current version of the platform is designed to work with TreeTagger[1] and ANNIE named entity recognition tool[2].
- `terminoFormDir`: Contains the terminological forms that are created using TERMINAE and output by it.
- `linguae`: Contains the search patterns that have been designed and their results (no pattern design tool is available in the current version).
- `thesauri`: Contains the termino-conceptual resources that are created using TERMINAE and output by it.
- `system`: Contains some files automatically created by TERMINAE.
- `yatea`: Contains the results of term extraction tools. The current version of the platform is designed to work with YaTeA term extractor[3].

3.2.2 How to import a project

A project to be imported is represented as a zipped file containing the project directory with all the required subdirectories and files of a given project. **You do not have to unzip the file**.

- Go to the main menu
- Click on `Terminae project actions`
- Click on `Import project`
- A first dialog window appears in which you must indicate the zipped file to load.

[1] http://www.ims.uni-stuttgart.fr/projekte/corplex/TreeTagger/

[2] http://gate.ac.uk/ie/annie.html

[3] http://search.cpan.org/%7Ethhamon/Lingua-YaTeA-0.5/

- A second dialogue window appears, to propose the directory into which the project will be imported. If you do not accept, you'll be offered to choose another one.

When the project is imported, its main characteristics are presented in the `Terminae project information` view on the left (by default) of the project perspective and you can start working on it.

3.2.3 How to create a project

To start working on a new project:

- Go to the main menu
- Click on `Terminae project actions`
- Click on `Create Terminae project`
- A first dialog window appears, in which you must indicate the name of the project.
- A second dialogue window appears, in which you must indicate in which directory you want to locate the project. A directory with the same name as the project is automatically created with 6 subdirectories.

To start working on your project to build termino-ontological resource from a given corpus, you need to have at least the following files in your project directory (more details in 6.1.1):

- In the `corpora` subdirectory:
  - A tagged version of the *row corpus(.txt)* (.tt file as output by TreeTagger)
- In the `yatea` subdirectory: the list of terms that have been extracted from the tagged version of the corpus by YaTeA (.xml file).

You must also give the name of the corpus if you exploit one and the name(s) of the authors(s) of the future resource(s).

When the project is created, its main characteristics are presented in the `Terminae project information` view on the left (by default) of the project perspective and you can start working on it.

3.3 Hidden files

The software creates 2 hidden files to manage the Terminae application:

- The file `.Terminae` contains the name of the current project. It is created in the directory where you launch the Terminae application. You normaly do not need to modify it.

- The file `.nameOfProject.xcfg` defines the configuration of each project (the set of files exploited by the project). Advertised user may easily understand its content, and may happen to change it in tricky cases (e.g. for renaming directories or files).

These files are text files or modifiable xml files.

Chapter 4

Main menu

Figure 4.1 presents the main menu of the TERMINAE platform, which is accessible from any perspective. It presents 4 items which are associated to specific actions or sub-menus[1]:

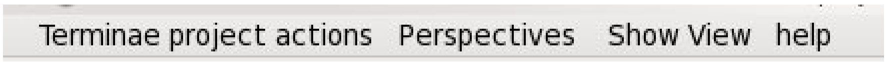


Figure 4.1: Main menu

- The action submenu gives access to the specific functionalities accessible at the Terminae level where you are currently working. The name of these action menu depends to the perspective from which it depends: `Terminae project actions`, `Linguistics actions`, `Terminological actions`, `TerminoConceptual actions` and `Terminae links`.
- The `Perspectives` item allows to open new perspectives: you simply have to click on the name of the perspective you want to open in the perspective list that appears. 5 perspectives are accessible:
  - `Terminae Project perspective`, which is the default perspective which is opened when a project is loaded. It is presented in Section 5.
  - `Terminae Terminological level (step 1)` perspective (see Section 6).
  - `Terminae Terminological level (step 2)` perspective (see Section 7).

[1] This main menu slightly differe from on exploitation system to another.

  - `Terminae TerminoConceptual level` perspective (see Section 8).
  - `Neon toolkit Conceptual level (OWL)` perspective (see Section 9).

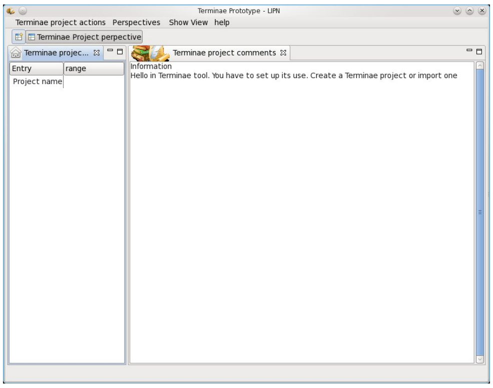
  The 4 first perspectives make up Terminae. The OWL perpective belongs to Neon ToolKit 2.4. Please note that the last Eclipse perspective (`Team Synchronizing`) is used by Neon ToolKit.

- The item `Search` is proposed Neon toolkit Conceptual level (OWL) perspective (it is not described in this report).
- The item `Help` is proposed in all eclipse application (it is not described in this report).
- An additional `Terminae` submenu is proposed on MacOS systems. It gives access to the standard application main operations: information (`About Terminae`), `Preferences`, `Hide Terminae`, `Quit Terminae`.

Chapter 5

Project management perspective

TERMINAE starts with the project management perspective. This perspective has 2 views (Fig. 5.1):


Terminae Prototype - LIPN
Terminae project actions
Perspectives
Show View
help
Terminae Project perpective
Terminae projec...
Terminae project comments

Entry	range
Project name	

Information
Hello in Terminae tool. You have to set up its use. Create a Terminae project or import one

Figure 5.1: Project management perspective

- The left view presents the project information if a project has been already defined: project, corpus, thesaurus and author(s)' names.
- The right view is a text editor where the user may write comments. To save the comments, you have to click on **`ctrl+s`**.

5.1 `Terminae project actions` menu

A project consists of all data used or created by TERMINAE when building a specific termino-ontological resource from a given corpus (see Section 3.2.1 for a description of the project structure).

The corpus is in a txt file (it is advised to use utf-8 encoding). See section 6.1.1 to have the description of the used files.

You can either:

- Create a new project (`Create Terminae project`) if you start to build a specific termino-ontological resource from a given corpus. You have to specify:
  - The name of your project.
  - The name of the directory where you want to locate your project. A default directory is proposed but click on the `cancel` button and navigate through the file system if you want to choose another directory.
- Switch from one project to another (`Load Terminae project`, note that only one project can be opened at the same time). You are first offered to navigate through the file system to select the directory containing the concerned project directory.
- Export the current project (`Export project`). A zipped file is created in which all the required directories and files are included. If you have created a Neon project, its directory is also included in the zipped file.
- Import an existing project (`Import project`). The project to be imported is represented as a zipped file containing the project directory with all the required subdirectories and files. You do not have to unzip the file but you have to specify:
  - The zipped file to load.
  - The name of the directory where you want the project to be imported.

5.2 `Help` menu

The Help information is not available yet.

5.3 `Show View` menu

Each perspective has many views and a main view which is on the left side of the perspective. A click on an item in the main view change values in other views. These views may be closed by the user or he/she may want to see a view of another perspective which is not in the used perspective (only one perspective may be selected).

This menu is used to reopen a view that has previously been closed. Click on the single item (`Other...`) to visualise the list of available views and choose again `Other` to find TERMINAE views. Select the view you want to reopen or to see - and be aware that the view may be dependant of one or the other perspective

Chapter 6

Terminae Terminological level (step 1) perspective

The Terminae Terminological level allows to browse and modify the list of domain specific lexical units that have been extracted from the source corpus using term extraction and named entity recognition tools such as YaTeA[1] and ANNIE[2].

TERMINAE assumes that the acquisition corpus has been processed by TreeTagger and YaTeA and possibly ANNIE beforehand. YaTeA takes as input:

- A tagged corpus (required),
- A list of terms extracted from it as input (required, see Section 6.1.1),
- Lists of named entities and named entity types (optional, see Section 6.1.2).

6.1 Data: Terminological files

6.1.1 Term files

When you open the `Terminological level perspective`, you have to specify the terminological data you want to start with (note that additional data can be loaded afterwards). You have to:

- Load a term list (`Load Yatea file`), which is supposed to be located in the `yatea` subdirectory of your project.
- Indicate how many documents your corpus encompasses. Note that documents are numbered starting from 1 if there are several of them but that a single document has number 0.

[1] http://search.cpan.org/%7Ethhamon/Lingua-YaTeA-0.5/

[2] http://gate.ac.uk/ie/annie.html

- Select the tagged corpus from which the terms have been extracted (`.tt` file). It is supposed to be located in the `corpus` subdirectory of your project.
- Speficy the corpus language: English (en) or French (fr).

When the terminological data is loaded, TERMINAE creates two additional files in the `yatea` directory:

- `fTempCorpus2XML.xml` which is an xml version of the corpus.
- `fTempTT2XML.xml` which is an xml version of the tagged corpus.

If you have several documents, each one must be processed by TreeTagger and the results must be concatenated in a single file where the various intial documents are separated by a document tag as shown below:

`Text_n TAB Document TAB n` where TAB is the tabulation character and n varies between 0 and x-1 (x being the total number of documents).

6.1.2 Named entity files

You may also want to work with named entities. In that case, you need two files that are output by the ANNIE named entity recognition tool (see Annex 10.7 for details on the file format) and which are expected to be located in the `corpus` subdirectory of your project:

- The first xml file indicates which named entity types you are interested in.
- The second xml file contains the list of named entities extracted by ANNIE.

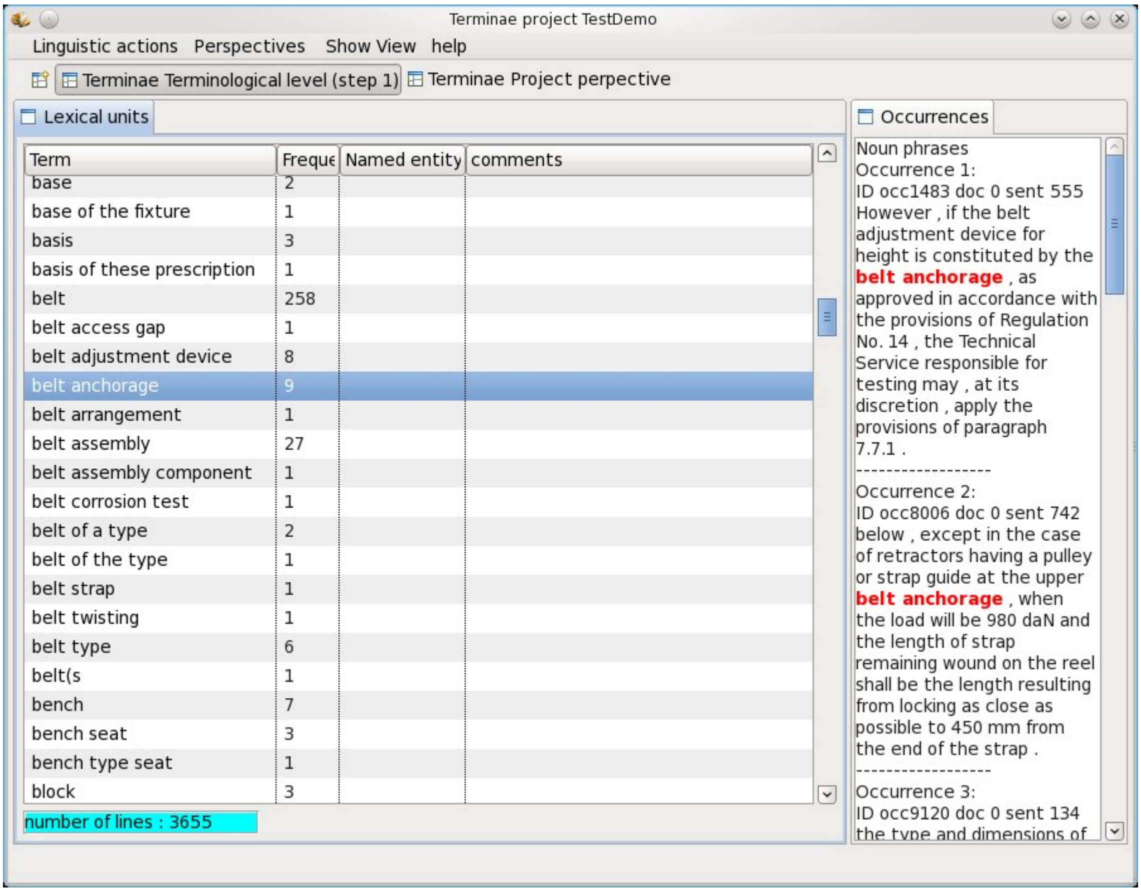
To create such files, follow the procedure described in Annex 10.8.

6.2 Perspective overview

If everything works properly when loading the terminological data, the window of Figure 6.1 appears when the `Terminae Terminological level (step 1)` perspective is first opened.

The window is composed of two views: the `Lexical units` view on the left and the `Occurrences` view on the right.

The terminological units (either terms or named entities) are listed on the left view. By clicking on the heads of the columns, you can sort the list alphabetically


Terminae project TestDemo
Linguistic actions
Perspectives
Show View
help
Terminae Terminological level (step 1)
Terminae Project perpective
Lexical units
Occurrences
number of lines : 3655

Term	Freque	Named entity	comments
base	2		
base of the fixture	1		
basis	3		
basis of these prescription	1		
belt	258		
belt access gap	1		
belt adjustment device	8		
belt anchorage	9		
belt arrangement	1		
belt assembly	27		
belt assembly component	1		
belt corrosion test	1		
belt of a type	2		
belt of the type	1		
belt strap	1		
belt twisting	1		
belt type	6		
belt(s	1		
bench	7		
bench seat	3		
bench type seat	1		
block	3		

Noun phrases
Occurrence 1:
ID occ1483 doc 0 sent 555
However , if the belt adjustment device for height is constituted by the **belt anchorage** , as approved in accordance with the provisions of Regulation No. 14 , the Technical Service responsible for testing may , at its discretion , apply the provisions of paragraph 7.7.1 .

Occurrence 2:
ID occ8006 doc 0 sent 742
below , except in the case of retractors having a pulley or strap guide at the upper **belt anchorage** , when the load will be 980 daN and the length of strap remaining wound on the reel shall be the length resulting from locking as close as possible to 450 mm from the end of the strap .

Occurrence 3:
ID occ9120 doc 0 sent 134
the type and dimensions of

Figure 6.1: Visualisation of Yatea results

(`Term`), by frequency (`Frequency`) or by type (terms vs. named entities) and named entity type (`Named entity`).

The last column of the `Lexical units` view allows to write comments: if you click on a cell comment, a text field appears and you can add a comment to the corresponding terminological unit. The comments are saved with the terminological results and can be reloaded upon request when the YaTeA results are loaded.

The occurrences of the selected terminological unit in the working corpus appear on the right view.

6.3 Linguistic actions menu

The action menu associated with the Terminae Terminological level (step 1) perspective is the Linguistic action menu. It proposes 3 submenus and 2 actions, which are also contextually accessible from the right click of the mouse:

- File submenu
- Term management submenu
- Cleaning submenu
- New terminological form action
- To terminologcial form action

Those submenus and actions are presented in the following subsections.

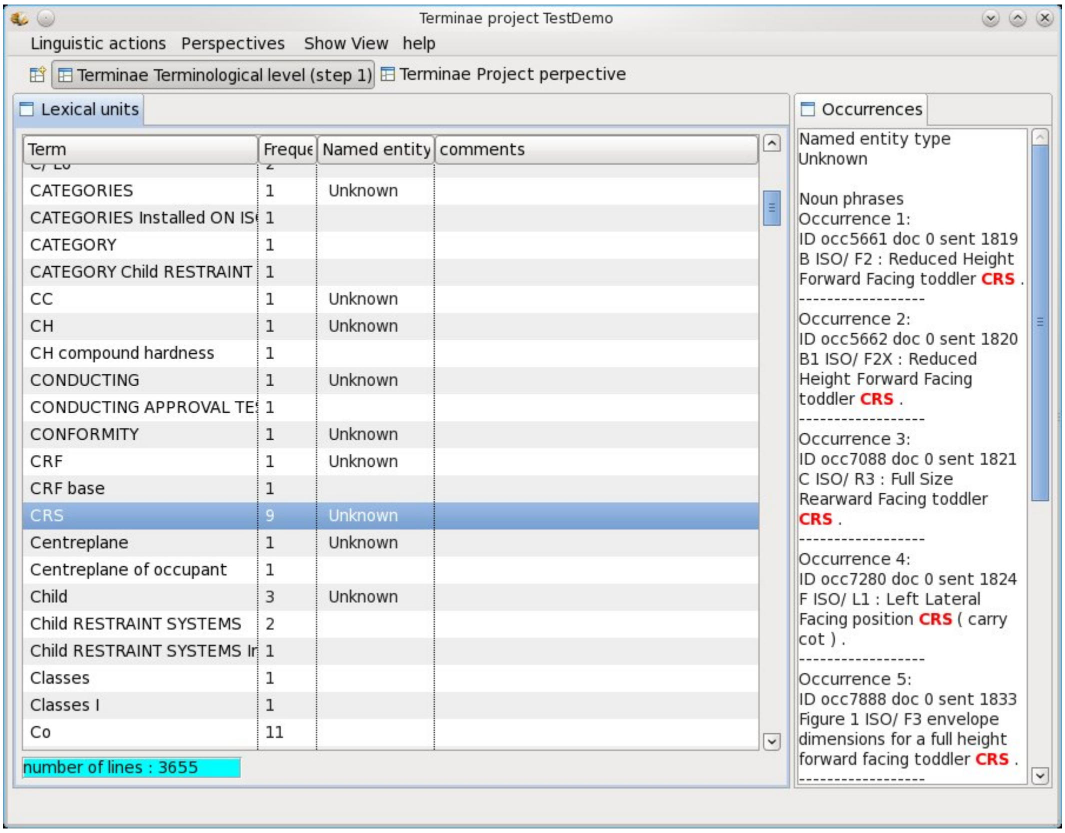
6.3.1 File submenu

This menu allows to load and save terminological data. It proposes the following actions:

- Load Yatea results to load the terms initially extracted from your corpus by YaTeA or saved in a XML backup. The procedure is the same as that described in Section 6.1.1.
- Save Yatea results to make an XML backup (see Annex 10.1 for details on the file format).
- Load named entities from ANNIE results to load the named entities identified by the ANNIE named entity recognition tool (see Section 6.1.2):
  - A first file dialog window opens, in which you have to indicate which named entity types you are interested in by selecting a named entity type XML file that should be located in the corpus subdirectory of your project.
  - A second file dialog window opens, in which you have to select another xml file containing the list of named entities extracted by ANNIE. This file should also be located in the corpus subdirectory of your project.

- `Save named entities` to make an XML backup (see Annex 10.2 for details on the file format).
- `Load named entities` to load the named entities from an XML backup.
- `Load all lexical units` to load the terms and named entities from a single XML backup.
- `Save all lexical units` to make an XML backup of all entities (terms and named entities) (see Annex 10.3 for details on the file format).

If everything works properly, when all types of terminological data are loaded, the window of Figure 6.2 appears.


Terminae project TestDemo
Linguistic actions
Perspectives
Show View
help
Terminae Terminological level (step 1)
Terminae Project perpective
Lexical units
Occurrences

Term	Freque	Named entity	comments
C/ L0	2		
CATEGORIES	1	Unknown	
CATEGORIES Installed ON IS	1		
CATEGORY	1		
CATEGORY Child RESTRAINT	1		
CC	1	Unknown	
CH	1	Unknown	
CH compound hardness	1		
CONDUCTING	1	Unknown	
CONDUCTING APPROVAL TE	1		
CONFORMITY	1	Unknown	
CRF	1	Unknown	
CRF base	1		
CRS	9	Unknown	
Centreplane	1	Unknown	
Centreplane of occupant	1		
Child	3	Unknown	
Child RESTRAINT SYSTEMS	2		
Child RESTRAINT SYSTEMS Ir	1		
Classes	1		
Classes I	1		
Co	11		

number of lines : 3655

Named entity type
Unknown

Noun phrases
Occurrence 1:
ID occ5661 doc 0 sent 1819
B ISO/ F2 : Reduced Height Forward Facing toddler **CRS** .

Occurrence 2:
ID occ5662 doc 0 sent 1820
B1 ISO/ F2X : Reduced Height Forward Facing toddler **CRS** .

Occurrence 3:
ID occ7088 doc 0 sent 1821
C ISO/ R3 : Full Size Rearward Facing toddler **CRS** .

Occurrence 4:
ID occ7280 doc 0 sent 1824
F ISO/ L1 : Left Lateral Facing position **CRS** (carry cot) .

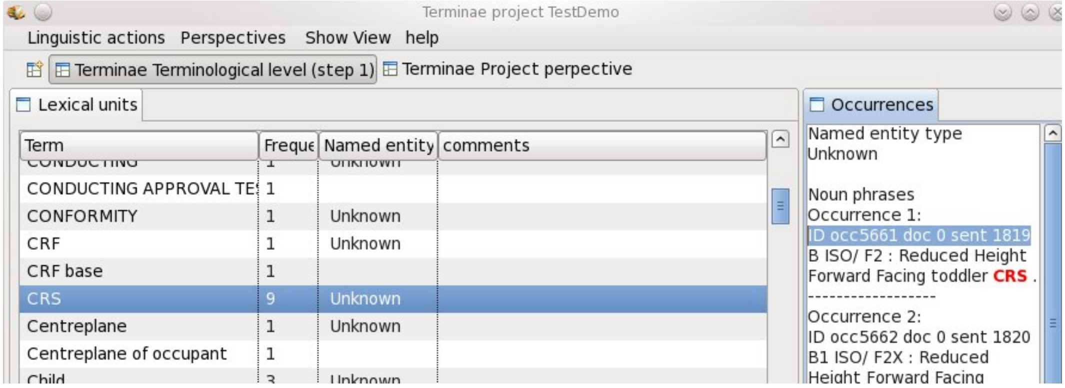
Occurrence 5:
ID occ7888 doc 0 sent 1833
Figure 1 ISO/ F3 envelope dimensions for a full height forward facing toddler **CRS** .

Figure 6.2: Visualisation of terms and named entities

6.3.2 `Term Management` submenu

This menu allows to manage terminological data, *i.e.* to visualise the list of terminological units and edit it by clustering, removing or adding some of them. The `Term Management` menu proposes 9 different actions:

- `Visualize all terms` to redisplay the list of terminological units after a search sequence.
- `Find a term` to search for a specific unit, on the basis of its beginning characters. Note that this functionality is also directly accessible when the list of terms is selected by typing the first letter of the searched term.
- `Cluster terms` to cluster several lexical units. You first have to select the various units you want to cluster, then click on the `Cluster terms` action and choose the canonical form you want to keep. The alternative forms are removed from the term list and all their occurrences are attached to the canonical form, which frequency count is updated.
- `Add a term` to add a new term to the term list.
- `Remove a term` to remove the selected term from the list
- `Undo remove` to undo the last remove action. This may also undo a cleaning action (see Section 6.3.3).
- `View occurrence context` to visualise the surrounding sentences of an occurrence. You have to select the occurrence identifier (see Figure 6.3) and to set the size of the expected context (expressed as a number of sentences).


Terminae project TestDemo
Linguistic actions
Perspectives
Show View
help
Terminae Terminological level (step 1)
Terminae Project perpective
Lexical units
Occurrences

Term	Freque	Named entity	comments
CONDUCTING	1	Unknown	
CONDUCTING APPROVAL TE	1		
CONFORMITY	1	Unknown	
CRF	1	Unknown	
CRF base	1		
CRS	9	Unknown	
Centreplane	1	Unknown	
Centreplane of occupant	1		
Child	3	Unknown	

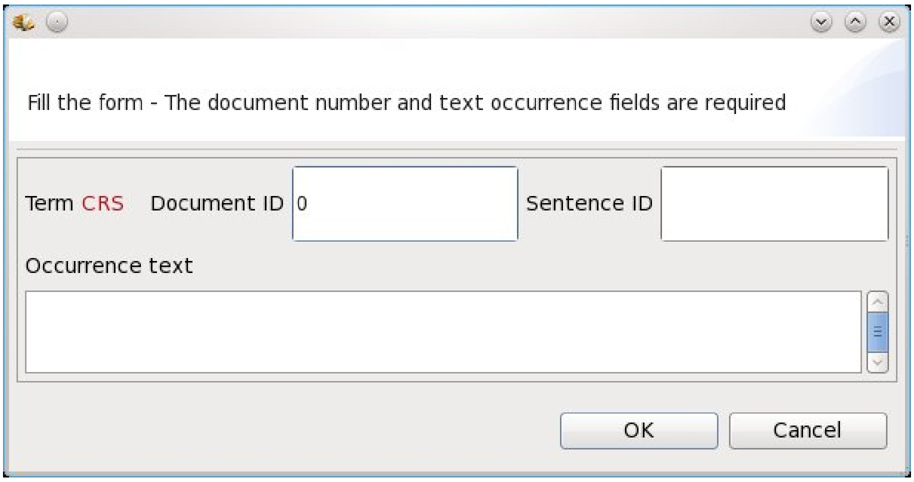
Named entity type
Unknown

Noun phrases
Occurrence 1:
ID occ5661 doc 0 sent 1819
B ISO/ F2 : Reduced Height
Forward Facing toddler **CRS** .

Occurrence 2:
ID occ5662 doc 0 sent 1820
B1 ISO/ F2X : Reduced
Height Forward Facing

Figure 6.3: Select an occurrence identifier

- `Add occurrence for a term` to enter a new occurrence for a term. You have to select a term and fill the form (see Figure 6.4).
- `Remove occurrence for a term` to remove an occurrence for a term. You have to select the identifier of the occurrence you want to remove.



Fill the form - The document number and text occurrence fields are required

Term CRS Document ID 0 Sentence ID

Occurrence text

Figure 6.4: Add occurrence for a term

6.3.3 `Cleaning` submenu

This menu allows to clean up the list of terminological units by removing a certain category of terms or named entities. Various options are proposed:

- `Remove terms listed in a file` allows to suppress all the terminological units that are listed in a given file. You have to give the name of that file, in which the stop words are listed, one at each line.
- `Remove terms involving given characters` allows to clean the list of terminological units on a character basis. You have to type in the list of forbidden characters.
- `Remove single-character terms` allows to suppress the single-character terms from the list of terminological units.
- `Remove adjectives` allows to suppress the terms that are tagged as adjectives.
- `Remove numbers` allows to suppress the terms that are numbers.

- `Remove adverbs` allows to suppress the terms that are tagged as adverbs.

6.3.4 Terminological form actions

This menu is used to define terminological forms described in next chapter.

- `New terminological form` allows to create a terminological form for the selected term. Once the terminological form is created, the new form can be visualized on the `Terminae Terminological level (step 2)` perspective, which is automatically opened, and the lexical unit which form has been created is displayed in blue character in the `Lexical units` view(`Terminae Terminological level (step 1)` perspective.

- `To terminological form` allows to visualise the terminological form of the selected terminological unit if it has one. This action automatically switches from the `Terminae Terminological level (step 1)` perspective to the `Terminae Terminological level (step 2)` perspective.

Chapter 7

Terminae Terminological level (step 2) perspective

This perspective can be opened either by creating a terminological form or from the main Perspective menu (`Terminological level (step 2)`).

7.1 Perspective overview

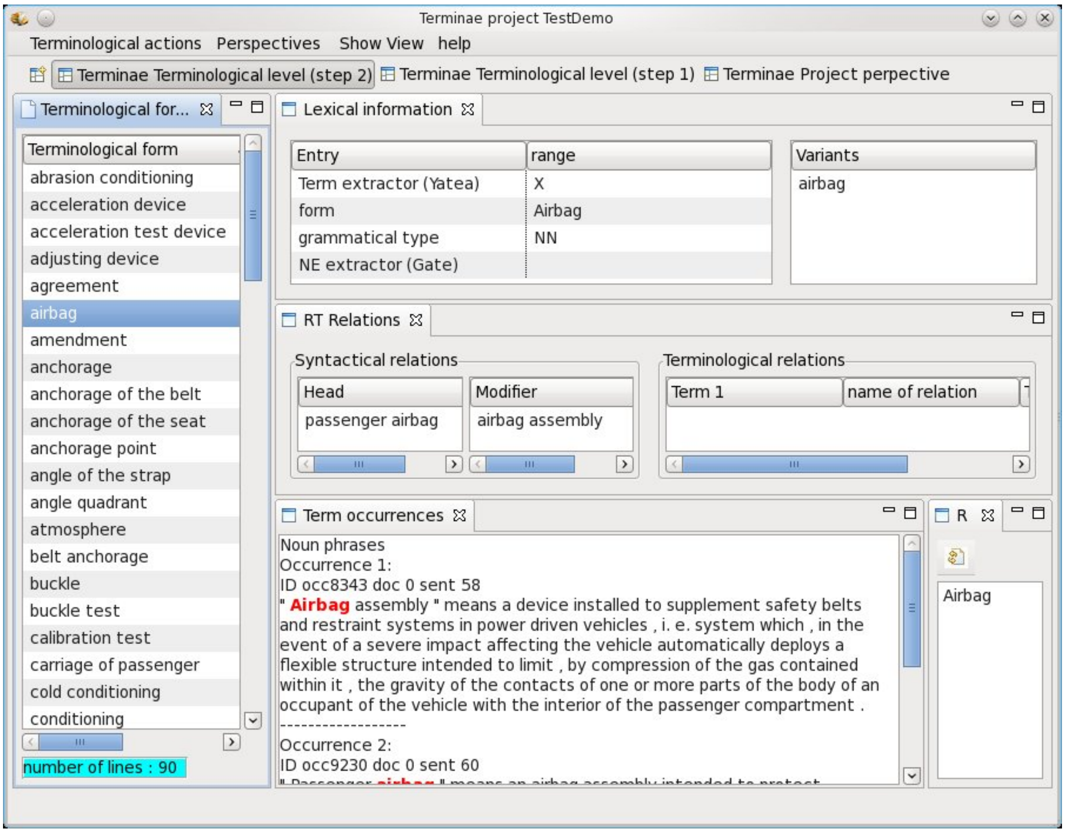
The `Terminae Terminological level (step 2)` perspective is composed of two main parts, with a global view on the left and a set of more detailed and dependant views on the right (see Figure 7.1):

- The `Terminological form list` view is by default presented on the left part of the perspective. It gives the lists of all the canonical terminological units for which a terminological form has been created (the form can be `In progress` or `Completed`).
- The other views form the terminological form of the unit that has been selected in the `Terminological form list` (see Section 7.2).

Note that, when the list of terminological forms is selected, you can find any terminological form by typing the first letter of its canonical terminological unit.

7.2 Data: Terminological forms

An example of terminological form is displayed on the right part of Figure 7.1. A terminological form gathers all the lexical and terminological information that has been collected or manually added for a given term or named entity. It is usually composed of the following views:


Terminae project TestDemo
Terminological actions
Perspectives
Show View
help
Terminae Terminological level (step 2)
Terminae Terminological level (step 1)
Terminae Project perpective
Terminological for...
Lexical information
RT Relations
Term occurrences
R
Airbag

Terminological form
abrasion conditioning
acceleration device
acceleration test device
adjusting device
agreement
airbag
amendment
anchorage
anchorage of the belt
anchorage of the seat
anchorage point
angle of the strap
angle quadrant
atmosphere
belt anchorage
buckle
buckle test
calibration test
carriage of passenger
cold conditioning
conditioning

number of lines : 90

Entry	range
Term extractor (Yatea)	X
form	Airbag
grammatical type	NN
NE extractor (Gate)	

Variants
airbag

Syntactical relations

Head	Modifier
passenger airbag	airbag assembly

Terminological relations

Term 1	name of relation

```
Noun phrases
Occurrence 1:
ID occ8343 doc 0 sent 58
" Airbag assembly " means a device installed to supplement safety belts
and restraint systems in power driven vehicles , i. e. system which , in the
event of a severe impact affecting the vehicle automatically deploys a
flexible structure intended to limit , by compression of the gas contained
within it , the gravity of the contacts of one or more parts of the body of an
occupant of the vehicle with the interior of the passenger compartment .
------------------
Occurrence 2:
ID occ9230 doc 0 sent 60
" Passenger airbag " means an airbag assembly intended to protect
```

Figure 7.1: `Terminae Terminological level (step 2)` perspective

- The `Lexical information` view is a form in which you can freely create, modify or suppress some fields. By default four lexical fields are defined:
  - `Term extractor (Yatea)`, which range is X if the terminological unit has been extracted by YaTeA term extractor,
  - `form`, which gives its canonical form,
  - `grammatical type`, which gives its grammatical category;
  - `NE extractor (Gate)`, which the range is its type if the lexical unit is a recognised named entity.

  The first three fields are automatically filled in by information provided by YaTeA. The last one is an ANNIE (Gate) information.

- The `Variants` view lists all the lexical forms that are associated as variants to the canonical form. They can be found in the corpus or manually added.

- The `Relations` view presents the relations that the terminological unit has:
  - The `Syntactical relations` list shows the phrases to which it belongs either as a head or as a modifier. The syntactical information is provided by YaTeA analysis of the corpus.
  - The `Terminological relations` list shows what are its terminological relationships. In the current version of the TERMINAE platform, the terminological relations have to be filled manually.
- The `Occurrences` view lists all the occurences of the terminological unit that have been identified. They can be occurrences of the canonical form or of any of its alternative (variant) form.
- The `Related termino-concepts` view shows to which termino-concepts the terminological unit is related.

As indicated in the second column of the `Terminological form list` view, a terminological form can be `In progress` or `Completed`.

Each terminological form is saved in an XML file in the `terminoFormDir` directory. The list of terminological forms is saved in the file `tableTermeFiches.xml` in `terminoFormDir` directory.

7.3 `Terminological actions` menu

The action menu associated with the `Terminae Terminological level (step 2)` perspective is the `Terminological action` menu. It proposes 3 submenus which are presented in the following subsections:

- `Termino-concept management` submenu
- `Form management` submenu
- `Feature management` submenu

The corresponding actions are also contextually accessible from the right click of the mouse.

7.3.1 `Termino-concept management` submenu

This submenu proposes three different actions:

- `Create a termino-concept` to create a termino-concept linked to the selected terminological unit. The termino-concept is added to the current thesaurus. If the terminological unit is a named entity, the type of the named entity may also give bearth to a termino-concept and a kindOf link is created between the two termino-concepts.
- `Remove a termino-concept` to remove a termino-concept from the current thesaurus.
- `To TerminoConceptual level` to switch from the `Terminae Terminological level (step 2)` perspective to the `Terminae TerminoConceptual level` perspective.

7.3.2 `Form management` submenu

This submenu proposes two actions related to terminological forms:

- `Remove a terminological form`
- `Validate a terminological form`: this action is used to note that the work on this terminological form is completed. It acts as a comment aimed at the user.

7.3.3 `Feature management` submenu

This submenu proposes various actions related to the detailed information provided for a given terminological unit and recorded in its terminological form:

- `Add a variant` to add a lexical variant of the selected term.
- `Remove a variant` to remove a lexical variant of the selected term.
- `Add a lexical entry` to add a lexical entry for the selected term. You have to type in the entry name and its value separated by two points.
- `Remove a lexical entry` to remove a lexical entry.
- `Add a syntactical relation-head` to add a phrase where the selected term is the head.
- `Add a syntactical relation-modifier` to add a phrase with the selected term as a modifier.
- `Remove a syntactical relation-head` to remove the selected relation.

- `Remove a syntactical relation-modifier` to remove the selected relation.

- `Add a terminological relation` to add a terminological relation where the selected term is `term1` or `term2`,

- `Remove a terminological relation` to remove a terminological relation.

- `Add an occurrence` to add an occurrence to the selected term. You have to specify the document identifier and to type in the text of the occurrence.

- `Remove an occurrence` to remove an occurrence to the selected term. Select the relevant occurrence to indicate which occurrence has to be removed.

Chapter 8

Terminae TerminoConceptual level perspective

This perspective must be opened from the `Perspective` submenu in the main menu by selecting the `Terminae TerminoConceptual level`.

8.1 Perspective overview

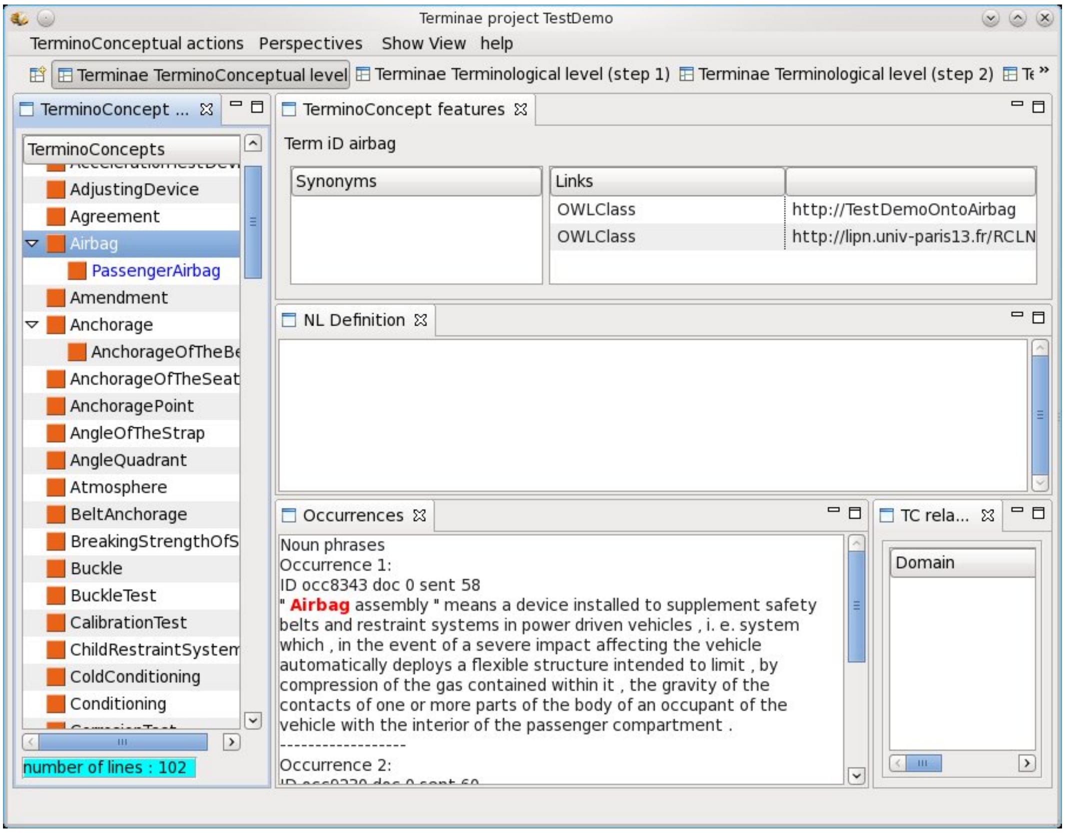
The `Terminae TerminoConceptual level` perspective presentation is very similar to that of the `Terminae Terminological level (step 2)` perspective. It is composed of two main parts, with a global view on the left and a set of more detailed and dependant views on the right (see Figure 8.1):

- The `TerminoConcept tree` view is, by default, presented on the left part of the perspective. It shows the hierarchy of all the termino-concepts that have been created.
- The other views form the termino-conceptual form of the termino-concept that has been selected in the `TerminoConcept tree` (see Section 8.2).

Note that you can find a termino-concept simply by typing its first letter in the `TerminoConcept tree` view.

8.2 Data: Termino-conceptual forms

The termino-conceptual level is a bridge between the terminological level and the conceptual level (the ontology). It is made of a set of termino-concepts which are themselves described by termino-conceptual forms gathering the relevant information that has been collected or defined for those termino-concepts.



Term iD airbag

Links	
OWLClass	http://TestDemoOntoAirbag
OWLClass	http://lipn.univ-paris13.fr/RCLN

Noun phrases
Occurrence 1:
ID occ8343 doc 0 sent 58
" **Airbag** assembly " means a device installed to supplement safety belts and restraint systems in power driven vehicles , i. e. system which , in the event of a severe impact affecting the vehicle automatically deploys a flexible structure intended to limit , by compression of the gas contained within it , the gravity of the contacts of one or more parts of the body of an occupant of the vehicle with the interior of the passenger compartment .

Occurrence 2:

Figure 8.1: `Terminae TerminoConceptual level` perspective

A termino-conceptual form is usually composed of the following views:

- The `TerminoConcept features` view presents the properties of the selected termino-concept:
  - its `Synonyms`,
  - its `Links`, that have been derived from the terminological levels. This mainly holds for termino-concepts related to named entities for which type information can be collected. Typical links are `brother`, `father` links.
- The `NL definition` view allows to enter a natural language definition for the selected termino-concept.
- The `Occurrences` view presents the occurrences in the corpus of the lexical units to which the termino-concept is linked.

- The `TC relations` view presents the termino-conceptual relations in which the termino-concept is domain or range.

Note that the meaning of a termino-concept is not formally defined. It is mainly described by its related occurrences.

8.3 `TerminoConceptual actions` menu

The action menu associated with the `Terminae TerminoConceptual level` perspective is the `TerminoConceptual action` menu. It proposes 4 submenus which are presented in the following subsections:

- `File` submenu
- `Termino-concept management` submenu
- `Feature management` submenu
- `Neon ontology` submenu

The corresponding actions are also contextually accessible from the right click of the mouse.

8.3.1 `File` submenu

This menu allows to load and save termino-conceptual data. It proposes the following actions:

- `Load XML format` to load a thesaurus in XML format (see DTD in Annex 10.4).
- `Save XML format` to save a thesaurus in XML format.
- `Import SKOS` to load an existing thesaurus in Skos format.
- `Export SKOS` to export a thesaurus in Skos format. A dialog window opens, in which you have to define an URI (added to the name of skos concepts to guarantee they are uniquely identified ; for instance http://www.lipn.univ-paris13.fr/terminae). Note that, in the current version of the TERMINAE platform, the termino-conceptual relations are not described in the exported file.
- `Export SKOS RDF/XML format` to export a thesaurus in RDF/XML format. A dialog window opens, in which you have to define an URI as for the skos format.

8.3.2 `Termino-concept management` submenu

- `Create termino-concept` to create a new termino-concept. You have to type in the name of the termino-concept if it is not created directly from a terminological unit.

- `Remove termino-concept` to remove the selected termino-concept. You have to confirm the removal.

- `Rename termino-concept` to change the name of the selected termino-concept.

- `Add kindOf link` to give a father to the selected termino-concept. A dialog window opens, in which you have to give the name of the father termino-concept.

- `Remove kindOf link` to remove a father of the selected termino-concept.

- `Add a RTC` to add a termino-concept relation for the selected termino-concept:

  - A first dialog window opens, in which you have to give the name of the relation.
  - A second dialog window opens, in which you have to click on `ok` if the selected termino-concept is the domain and on `cancel` if not.
  - A third dialog window opens, in which you have to give the name of the range or domain (depending on the previous answer). That termino-concept must pre-exist.
  - A choice dialog window then opens, in which you have to select the skos type of the relation.

- `Remove a RTC` to remove the selected termino-conceptual relation.

- `Add occurrence` to add an occurrence to the selected termino-concept.

- `Remove occurrence` to remove an occurrence of the selecteed termino-concept. You have to select the identifier of the occurrence to be removed.

- `Create a terminological form` to create a terminological form from a termino-concept. This functionality is useful when you want to add terminological information and occurrences to an existing thesaurus. You start from an existing termino-concept and create a terminological form using a defined corpus.

- `Create all terminological forms` to create all terminological forms from a preexisting thesaurus. This functionality is useful when you want to add terminological information and occurrences to an existing thesaurus. You start from an existing thesaurus and create a terminological form for each termino-concept using a defined corpus.

8.3.3 `Feature management` submenu

This submenu proposes various actions related to the detailed information provided for a given termino-concept and recorded in its termino-conceptual form:

- `Add a synonym` to add a synonym to the selected termino-concept. A dialog window opens for capturing the new synonym. If the corresponding terminological unit has been found by YaTeA or ANNIE, its occurrences are automatically clustered with that of the current termino-concept.

- `Remove a synonym` to remove a synonym. You have to confirm if you want also to remove the related occurrences.

- `Add a link` to add a type of link and its value.

- `Remove a link` to remove a type of link and its value.

8.3.4 `Neon ontology` submenu

This menu is used to link TERMINAE and Neon ToolKit. It supports the creation of the conceptual level and many actions to connect it to the termino-conceptual one:

- `Create a Neon project` is used to create a Neon toolkit project. If you want to work at the conceptual level, you have to create a Neon project and to specify its name. It is recommended to use different names for the TERMINAE and Neon projects.

- `Create Neon Toolkit ontology` is used to create an ontology. This ontology is part of the newly created Neon project.

- `Create a class` is used to create a class in the previous ontology and from the selected termino-concept. A dialog window opens, in which you have to give a name to the class and select a class father in the existing ontology. The class can be visualized in the `Neon toolkit Conceptual level (OWL)` perspective (see Figure 8.2). Note that the class is created with an annotation property in which the link to the source termino-concept and its identifier is saved. Once it has been linked to a class at the conceptual level, the termino-concept is displayed in blue color in the `TerminoConcept tree`.

- `To ontology level` is used to switch from the termino-conceptual perspective to the OWL one. This action opens the OWL perspective and shows the class corresponding to the selected termino-concept.

- `Link to Neon project` is used when one wants to exploit an existing Neon toolkit project.

- `Link to Neon ontology` is used when one wants to exploit an existing ontology in a specified project.

- `Link to a class` is used to link a termino-concept to an existing class.

- `Create an ObjectProperty` is used to create an objectProerty from a termino-conceptual relation. A dialog window opens and you have to enter the name of the property, the father object property, its domain and range. The objectProperty is created with an annotation property in which the name and type of the source termino-conceptual relation are saved.

- `Link a RT and an ObjectProperty` is used to link a termino-conceptual relation to an existing objectProperty.

- `Link a RT and a class` is used to link a termino-conceptual relation to a an existing class.

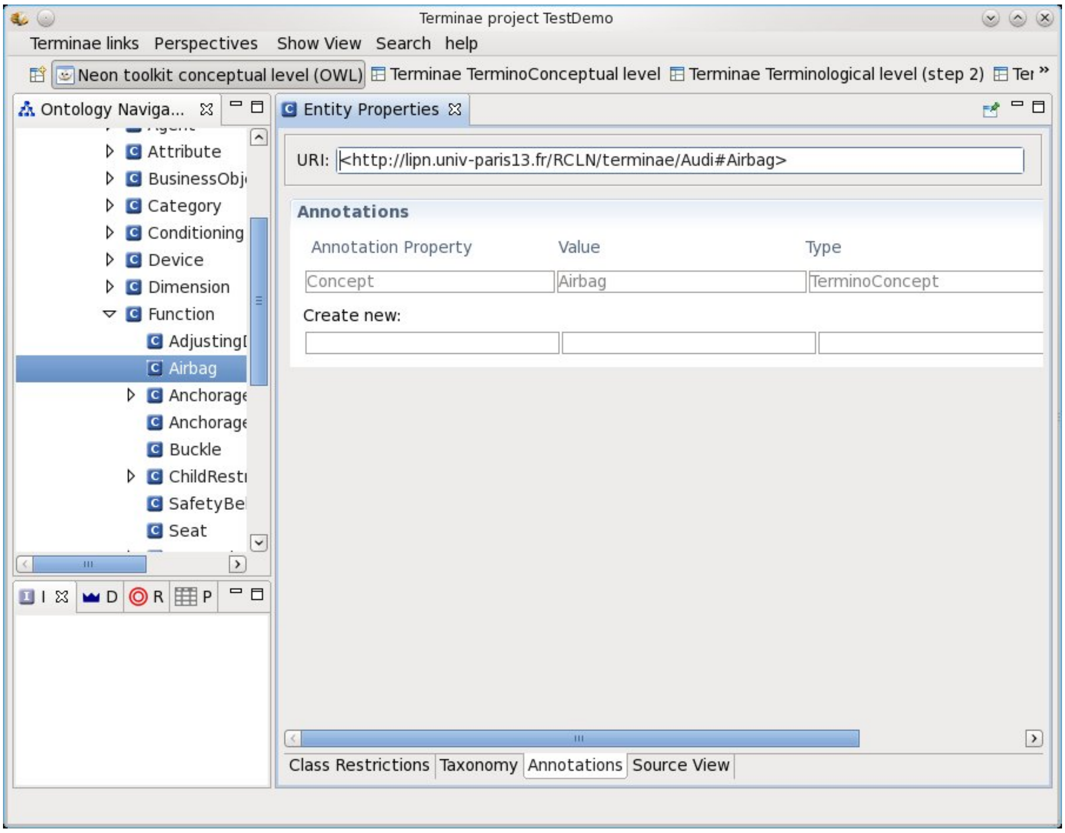
- `Create classes and TCs` is used to derive a set of classes from a set of selected termino-concepts. If these termino-concepts have termino-conceptual relations, objectProperties are created and linked to these source relations.

- `Create classes and TCs without dialog` offers the same functionality as above but there without dialog. The default values are systematically kept:

  - name of class = name of terminoconcept,
  - name of objectproperty = name of the RTC,
  - if termino-concepts are linked by a `isKindOf` link, the corresponding classes are in the same hierarchical order.

- `Link to an individual` is used to link a termino-concept to an individual. You have to enter the individual name and select the class from which it belongs thanks to dialog windows.

- `Create an individual` is used to create an individual. You have to enter the individual name and select the class from which it belongs thanks to dialog windows.



URI: <http://lipn.univ-paris13.fr/RCLN/terminae/Audi#Airbag>

Annotation Property	Value	Type
Concept	Airbag	TerminoConcept

Figure 8.2: `Neon toolkit conceptual level (OWL)` perspective

Chapter 9

Neon toolkit Conceptual level (OWL) perspective

The conceptual perpective is a Neon toolkit plugin (version 2.4) to which a specific menu has been added for the TERMINAE platform to link the conceptual and termino-conceptual levels.
When using Neon toolkit conceptual level perspective, you need to create or to import a Neon toolkit project (which is different from the Terminae project), and to create or import an ontology in this project.

This can be done either from theNeon ontology submenu of the Terminae TerminoConceptual perspective (Create a Neon project and Create Neon Toolkit ontology items) , or create the project and the ontology from the menu of the navigator view in Neon toolkit conceptual level perspective (click right).

In the Neon toolkit conceptual level perspective, you can also import an existing project. In this case, you have to refresh the view to display the imported project and to link it to the terminoConceptual perspective (see the following section). You can also import an ontology, use import item from the menu of the navigator view of Neon toolkit conceptual level perspective.

9.1 Perspective overview

The Neon toolkit Conceptual level (OWL) perspective presentation is very similar to that of the Terminae TerminoConceputal level perspective. It is composed of two main parts, with a global view on the left and a set of more detailed and dependant views on the right (see Figure 8.2). See the documentation (http://www.neon-toolkit.org/wiki/Documentation_and_Support).

9.2 Terminae links menu

Terminae links menu has been added to the Neon Toolkit perspective to link the conceptual and the termino-conceptual levels of Neon and TERMINAE projects and of the resulting termino-conceptual resources:

- To terminoConceptual level is used to switch from the Neon toolkit Conceptual level (OWL) perspective to the Terminae TerminoConceputal level perspective. Clicking on this action item (re-)opens the termino-conceptual perspective and selects the termino-concept associated with the class initially selected in the conceptual perspective.

- Create a termino-concept is used to create a termino-concept and link it to the selected. This functionality is useful when you want to add thesaurus information to an existing ontology. You start from an existing class and create a termino-concept in the thesaurus of the TERMINAE project.

- To link a class to a TC is used to link a class to an existing termino-concept in the thesaurus of the TERMINAE project.

Chapter 10

Annex

This annex lists the DTD used by Teminae.

10.1 XML backup DTD for terms

The DTD of the XML file which contains terms and their occurrences which is visualized in `Terminae Terminological level (step 1)` perspective.

```
<!ELEMENT DOC ( #PCDATA ) >
<!ELEMENT END_POSITION ( #PCDATA ) >
<!ELEMENT FORM ( #PCDATA ) >
<!ELEMENT ID ( #PCDATA ) >
<!ELEMENT LEMMA ( #PCDATA ) >
<!ELEMENT LIST_OCCURRENCES ( OCCURRENCE+ ) >
<!ELEMENT LIST_TERM_CANDIDATES ( TERM_CANDIDATE+ ) >
<!ELEMENT List_Variants ( Variant* ) >
<!ELEMENT MORPHOSYNTACTIC_FEATURES ( SYNTACTIC_CATEGORY ) >
<!ELEMENT NUMBER_OCCURRENCES ( #PCDATA ) >
<!ELEMENT OCCURRENCE ( ID, DOC, SENTENCE, START_POSITION,
                                 END_POSITION, Texte ) >
<!ELEMENT SENTENCE ( #PCDATA ) >
<!ELEMENT START_POSITION ( #PCDATA ) >
<!ELEMENT SYNTACTIC_CATEGORY ( #PCDATA ) >
<!ELEMENT TERM_CANDIDATE ( ID, LEMMA, FORM, List_Variants,
NUMBER_OCCURRENCES, LIST_OCCURRENCES, MORPHOSYNTACTIC_FEATURES ) >
<!ELEMENT TERM_EXTRACTION_RESULTS ( LIST_TERM_CANDIDATES ) >
<!ELEMENT Texte ( #PCDATA ) >
<!ELEMENT Variant ( #PCDATA ) >
```

10.2 XML backup DTD for ENs

The DTD of the XML file which contains named entities and their occurrences which is visualized in `Terminae Terminological level (step 1)` perspective.

```
<!ELEMENT DOC ( #PCDATA ) >
<!ELEMENT END_POSITION ( #PCDATA ) >
<!ELEMENT FORM EMPTY >
<!ELEMENT ID ( #PCDATA ) >
<!ELEMENT LEMMA ( #PCDATA ) >
<!ELEMENT LIST_EN ( NAMED_ENTITY+ ) >
<!ELEMENT LIST_OCCURRENCES ( OCCURRENCE* ) >
<!ELEMENT LIST_SENT ( SENT* ) >
<!ELEMENT List_Lemme EMPTY >
<!ELEMENT List_Variants EMPTY >
<!ELEMENT NAMED_ENTITY ( ID, LEMMA, FORM, List_Variants,
         Types, NUMBER_OCCURRENCES, LIST_OCCURRENCES, LIST_SENT ) >
<!ELEMENT NUMBER_OCCURRENCES ( #PCDATA ) >
<!ELEMENT OCCURRENCE ( ID, DOC, SENTENCE, START_POSITION,
                                        END_POSITION, Texte ) >
<!ELEMENT SENT ( ID, offset, phrase, List_Lemme ) >
<!ELEMENT SENTENCE ( #PCDATA ) >
<!ELEMENT START_POSITION ( #PCDATA ) >
<!ELEMENT Texte ( #PCDATA ) >
<!ELEMENT Types ( type+ ) >
<!ELEMENT offset ( #PCDATA ) >
<!ELEMENT phrase ( #PCDATA ) >
<!ELEMENT type ( #PCDATA ) >
```

10.3 EnsLexUnit DTD

The DTD of the XML file which contains terms, named entities and their occurrences which is visualized in `Terminae Terminological level (step 1)` perspective.

```
<!ELEMENT DOC ( #PCDATA ) >
<!ELEMENT END_POSITION ( #PCDATA ) >
<!ELEMENT Ens_Variants EMPTY >
<!ELEMENT FORM ( #PCDATA ) >
<!ELEMENT ID ( #PCDATA ) >
```

```
<!ELEMENT LEMMA ( #PCDATA ) >
<!ELEMENT LIST_EN ( NAMED_ENTITY+ ) >
<!ELEMENT LIST_OCCURRENCES ( OCCURRENCE* ) >
<!ELEMENT LIST_SENT ( SENT* ) >
<!ELEMENT LIST_TERM_CANDIDATES ( TERM_CANDIDATE+ ) >
<!ELEMENT List_Variants ( Variant* ) >
<!ELEMENT MORPHOSYNTACTIC_FEATURES ( SYNTACTIC_CATEGORY ) >
<!ELEMENT NAMED_ENTITY ( Ens_Variants | ID | LEMMA |
   LIST_OCCURRENCES |  LIST_SENT | NUMBER_OCCURRENCES | Types )* >
<!ELEMENT NUMBER_OCCURRENCES ( #PCDATA ) >
<!ELEMENT OCCURRENCE ( ID, DOC, SENTENCE, START_POSITION,
                                  END_POSITION, Texte ) >
<!ELEMENT SENT EMPTY >
<!ATTLIST SENT ID CDATA #REQUIRED >
<!ELEMENT SENTENCE ( #PCDATA ) >
<!ELEMENT START_POSITION ( #PCDATA ) >
<!ELEMENT SYNTACTIC_CATEGORY ( #PCDATA ) >
<!ELEMENT TERM_CANDIDATE ( ID, LEMMA, NUMBER_OCCURRENCES,
 LIST_OCCURRENCES, FORM, MORPHOSYNTACTIC_FEATURES, List_Variants,
            NAMED_ENTITY? ) >
<!ELEMENT TERM_EXTRACTION_RESULTS ( LIST_TERM_CANDIDATES, LIST_EN ) >
<!ELEMENT Texte ( #PCDATA ) >
<!ELEMENT Types ( type+ ) >
<!ELEMENT Variant ( #PCDATA ) >
<!ELEMENT type ( #PCDATA ) >
```

10.4 Thesaurus DTD

The DTD of the XML file which contains a thesaurus which is visualized in Terminae TerminoConceptual level perspective. A thesaurus contains a collection of terminoconcepts. Each terminoconcept is described by an ID, a natural language definition, corpus occurrences, a prefLabel, a set of “see_also”, a set of synonyms (altLabel), a set of children and its father.

```
<!ELEMENT DOC ( #PCDATA ) >
<!ELEMENT END_POSITION ( #PCDATA ) >
<!ELEMENT EnsTerminoConcepts ( name, TerminoConcept+ ) >
<!ELEMENT ID ( #PCDATA ) >
<!ELEMENT NL_Definition ( #PCDATA ) >
<!ELEMENT OCCURRENCE ( ID, DOC, SENTENCE, START_POSITION,
                     END_POSITION, Texte)>
<!ELEMENT PrefLabel ( #PCDATA ) >
<!ELEMENT RelationRTC ( name, domain, range, Skos_type ) >
<!ELEMENT SENTENCE ( #PCDATA ) >
<!ELEMENT START_POSITION ( #PCDATA ) >
<!ELEMENT See_also ( #PCDATA ) >
<!ELEMENT SetRTC ( RelationRTC? ) >
<!ELEMENT Skos_type ( #PCDATA ) >
<!ELEMENT Synonym ( #PCDATA ) >
<!ELEMENT TerminoConcept ( ID | NL_Definition | OCCURRENCE |
 PrefLabel | See_also | SetRTC | Synonym | children | fathers )* >
<!ELEMENT Texte ( #PCDATA ) >
<!ELEMENT child ( #PCDATA ) >
<!ELEMENT children ( child* ) >
<!ELEMENT domain ( #PCDATA ) >
<!ELEMENT father ( #PCDATA ) >
<!ELEMENT fathers ( father? ) >
<!ELEMENT name ( #PCDATA ) >
<!ELEMENT range ( #PCDATA ) >
```

10.5 TreeTagger English Tagset

```
CC Cooordinating conjunction
CD Cardinal number
DT Determiner
EX Existential there
FW Foreign word
IN Preposition or subordinating conjunction
JJ Adjective
JJR Adjective, comparative
JJS Adjective, superlative
LS list item marker
MD Modal
NN Noun, singular or mass
NNS Noun, plural
NP Proper noun, singular
NPS Proper noun, plural
PDT Predeterminer
POS Possessive ending
PP Personal pronoun
PP$ Possessive pronoun
```

```
RB Adverb
RBR Adverb, comparative
RBS Adverb, superlative
RP Particle
SYM Symbol
TO to
UH Interjection
VB Verb, base form
VBD Verb, past tense
VBG Verb, gerund or present participle
VBN Verb, past participle
VBP Verb, non-3rd person singular present
VBZ Verb, 3rd person singular present
WDT Wh-determiner
WP Wh-pronoun
WP$ Possesive wh-pronoun
WRB Wh-adverb
```

10.6 TreeTagger French Tagset

```
ABR abreviation
ADJ adjective
ADV adverb
DET:ART article
DET:POS possessive pronoun (ma, ta, ...)
INT interjection
KON conjunction
NAM proper name
NOM noun
NUM numeral
PRO pronoun
PRO:DEM demonstrative pronoun
PRO:IND indefinite pronoun
PRO:PER personal pronoun
PRO:POS possessive pronoun (mien, tien, ...)
PRO:REL relative pronoun
PRP preposition
PRP:det preposition plus article (au,du,aux,des)
PUN punctuation
PUN:cit punctuation citation
SENT sentence tag
SYM symbol
```

```
VER:cond verb conditional
VER:futu verb futur
VER:impe verb imperative
VER:impf verb imperfect
VER:infi verb infinitive
VER:pper verb past participle
VER:ppre verb present participle
VER:pres verb present
VER:simp verb simple past
VER:subi verb subjunctive imperfect
VER:subp verb subjunctive present
```

10.7 Use ANNIE to extract named entities

This annex describes the procedure to be followed to use ANNIE to extact named entities from a given document (only one document can be processed at a time).

Note that the following procedure is extracted from the Gate documentation for processing English corpora: http://gate.ac.uk/sale/tao/splitch3.html.

GATE enables you to extract named entities from plain texts and annotate your corpus with it. GATE is distributed with an IE system called ANNIE. ANNIE relies on finite state algorithms and the JAPE[1] language.

Take one large pile of text (documents, emails, etc.). Call this your corpus.

If you right-click on ”Language Resources” in the resources pane, select ”New” then ”GATE Document”, the window ”Parameters for the new GATE Document” will appear.

Once you indicate the corpus to work on it, you can call for ANNIE.

From the File menu, select ”Load ANNIE System”. To run it in its default state, choose ”with Defaults”. This will automatically load all the ANNIE resources, and create a corpus pipeline called ANNIE with the correct resources selected in the right order, and the default input and output annotation sets.

If ”without Defaults” is selected, the same processing resources will be loaded, but a popup window will appear for each resource, which enables the user to specify a name, location and other parameters for the resource. This is exactly the same procedure as for loading a processing resource individually, the difference being that the system automatically selects those resources contained within ANNIE. When the resources have been loaded, a corpus pipeline called ANNIE will be created as before.

[1] JAPE is a Java Annotation Patterns Engine. It provides finite state transduction over annotations based on regular expressions. JAPE allows you to recognise regular expressions in annotations on documents.

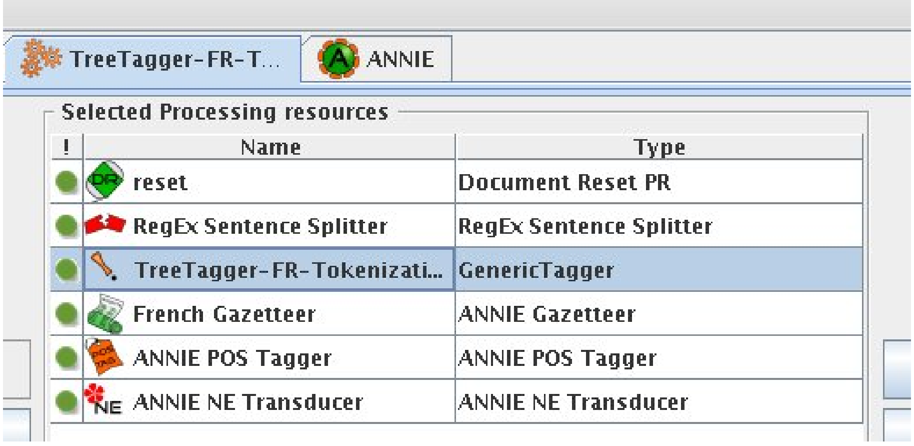
The next step is to add a corpus, and select this corpus from the drop-down corpus menu in the Serial Application editor. Finally click on ”Run” from the Serial Application editor, or by right clicking on the application name in the resources pane and selecting ”Run”.

To view the results, double click on one of the document contained in the corpus processed in the left hand tree view. No annotation sets nor annotations will be shown until annotations are selected in the annotation sets; the ”Default” set is indicated only with an unlabelled right-arrowhead which must be selected in order to make visible the available annotations. Open the default annotation set and select some of the annotations to see what the ANNIE application has done.

Having selected an annotation type in the annotation sets view, hovering over an annotation in the main resource viewer or right-clicking on it will bring up a popup box containing a list of the annotations associated with it, from which one can select an annotation to view in the annotation editor, or if there is only one, the annotation editor for that annotation.

Now to save your corpus annotated with ANNIE, right-click on a document in the resources tree and choose ”Save as XML”. In addition, all documents in a corpus can be saved as individual XML files into a directory by right-clicking on the corpus in the resources tree and choosing the option ”Save as XML”.

For French corpora, you have to install treetagger and load the Tagger_Framework plugin. In the resource directory, you find *TreeTagger-FR-Tokenization.gapp*. You load this application in Gate platform.You also load the Lang_French plugin and the french.gapp Gate application. The selected processing resources are defined in Figure 10.1 .



Selected Processing resources

!	Name	Type
●	reset	Document Reset PR
●	RegEx Sentence Splitter	RegEx Sentence Splitter
●	TreeTagger-FR-Tokenizati...	GenericTagger
●	French Gazetteer	ANNIE Gazetteer
●	ANNIE POS Tagger	ANNIE POS Tagger
●	ANNIE NE Transducer	ANNIE NE Transducer

Figure 10.1: Selected processing resources

10.8 Gate named entity type file

The DTD of the XML file which contains named entity type file which is used when loading named entities (see 6.1.2).

```
<?xml version=’1.0’ encoding=’UTF-8’?>
<ensTypeEn>
<typeEn>Organization</typeEn>
<typeEn>Date</typeEn>
<typeEn>Person</typeEn>
<typeEn>Percent</typeEn>
<typeEn>Location</typeEn>
<typeEn>Money</typeEn>
<typeEn>Title</typeEn>
<typeEn>Address</typeEn>
<typeEn>Unknown</typeEn>
<typeEn>Jobtitle</typeEn>
<typeEn>FirstPerson</typeEn>
<typeEn>Location</typeEn>
<typeEn>UrlPre</typeEn>
</ensTypeEn>
```